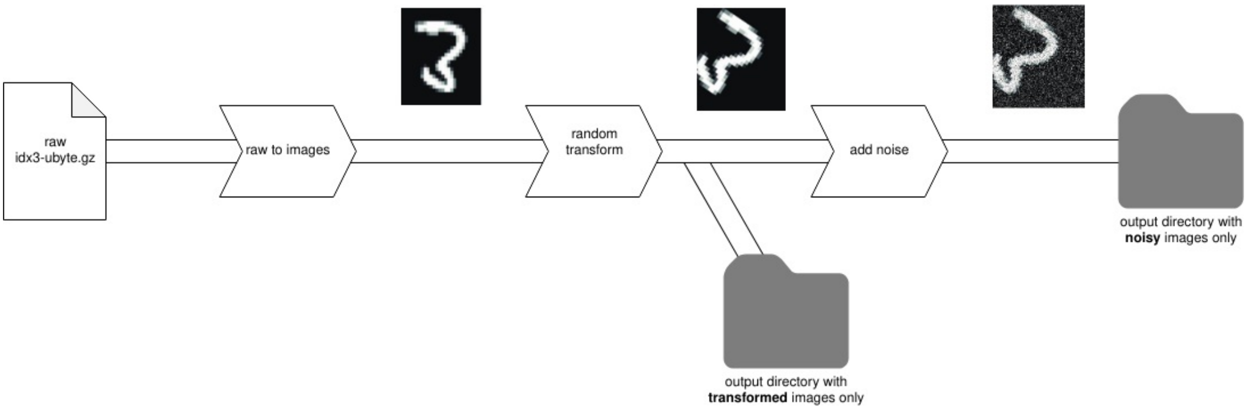
CCG Data Pipeline Challenge

Introduction

Your task is to design and develop a small data pipeline that inputs a raw dataset and outputs transformed noisy images. This pipeline is meant to be used in the future to train a denoising autoencoder (not to do in this challenge).

The pipeline has 3 modules:

- raw to images
- images to transformed images
- images to noisy images



Requirements

Multiprocessing

We expect you to use parallelization and/or concurrency to make the full pipeline as quick as you can.

Unittest

We expect you to provide a fair test coverage of the pipeline.

Containerization (optional)

You may use a containerization system for your pipeline such as Docker.

Tasks

Raw data to images

The first module must decompress the raw dataset (the file `train-images-idx3-ubyte.gz` present in this directory) to a set of images (either on disk or memory).

Images to transformed images

For each image in the dataset we want to create `N` new images where random transformations have been applied.

The transformations are:

- rotation from -100 to 100 degrees following a uniform distribution
- translation from -40% to 40% both axis following a uniform distribution

The transformed images must be saved in a directory.

Images to noisy images

For each transformed image, this module adds noise and saves the result in a different folder. We want a way to link the transformed image and the related noisy image together by the way you name them.